Exploring a Text Corpus via a Knowledge Graph*

Eleonora Bernasconi[1][0000-0003-3142-3084],
Miguel Ceriani[2][0000-0002-5074-2112], and
Massimo Mecella[1][0000-0002-9730-8882]

[1] Sapienza Università di Roma, ITA
{bernasconi,mecella}@diag.uniroma1.it
[2] Università degli Studi di Bari Aldo Moro, ITA
miguel.ceriani@uniba.it

Abstract. Semantic enrichment methods may be used to identify relevant entities in textual documents. These extracted entities are part of knowledge graphs and thus linked by semantic relationships. This work explores the idea of navigating the semantic relationships among extracted entities as a way to search a text corpus. A modular software system (including document management, semantic enrichment, data consolidation, and data integration) has been designed, to offer a visual user interface for such navigation on top of an arbitrary corpus of textual documents. The software, called ARCA, has been used in a real use case: to search in the book catalogue of a publishing house. The evaluation carried out with a set of potential users has shown so far the feasibility and effectiveness of the approach. Critical issues and potential limitations of the paradigm have also been found and are discussed.

Keywords: Semantic enrichment· Knowledge graph · Visual search interface

1 Introduction

Searching and exploring a vast text corpus has often arisen as a human need. Traditionally, a search is based on manually curated metadata classifying documents, for example, by arguments, authors, etc. In the digital era, the search moved from physical cabinets to databases. Albeit being a useful paradigm, the maintenance of metadata is an expensive process, which becomes progressively more expensive and less reliable with the increase of required detail. Efforts of

* This work has been partly supported by projects ARCA (POR FESR Lazio 2014–2020 - Avviso pubblico “Creatività 2020”, domanda prot. n. A0128-2017-17189) and STORYBOOK (POR FESR Lazio 2014-2020 - Avviso Pubblico “Progetti di Innovazione Digitale”, domanda prot. n. A0349-2020-34437).

transitioning to electronic documents (either created natively as such or digitized) have helped, enabling the direct text-based search of the content. Semantic enrichment methods, as *named-entity recognition and linking (NERL)* [14,18], aim at bridging the semantic gap between raw text and concepts, by associating words in the documents with entities in a knowledge base, often a knowledge graph (KG).

NERL successfully enabled users to search and analyze text corpora more effectively [17]. While knowledge extraction methods as NERL are now broadly used by big players in the industry as well as in academic projects, their usage by small to medium size organizations is still minimal.

1.1 Research questions

For the sake of the analytic approach, we frame our effort through a set of research questions. The questions elicited below are relevant to the application of KG-based approaches for the exploration of text corpora.

Q1. Would users, exploring a corpus of text, profit from the semantic navigation of the associated KG of topics?

Q2. What kind of user interface would effectively support such navigation?

Q3. What kind of users, scenarios, and tasks would benefit from this interaction paradigm?

Q4. Do building and maintaining a semantic enrichment and KG creation pipeline necessarily involve high upfront costs and highly skilled developers?

1.2 Hypotheses

To try to reply to the questions above, we designed the presented study.

H1 (relevant to Q1 and Q2). Users will be able to explore a text corpus through a KG-based user interface which offers the following main functions: *(a)* find concepts through text search, *(b)* visually navigate the concepts and their relationships, and *(c)* show documents relevant to the selected concept.

H2 (relevant to Q3). Given a corpus of text in a specific domain, it will benefit both users with little knowledge of the domain (by supporting semantically-relevant discovery) and domain experts (by enabling topic-oriented visual organizations of the documents).

H3 (relevant to Q4). It is feasible to build a ready-to-use complete system, including both semantic enrichment pipeline and web-based front end, able, with only some configuration, to be applied to any specific corpus to enable the KG-based exploration.

1.3 Approach

A comprehensive software system has been previously envisioned and proposed [8] to address the research questions. The system, which now has been fully implemented and evaluated on a specific use case, is meant to enable the KG-based exploration of a given text corpus and hence test our hypotheses. The system is organized according to the following main functions:

– extraction of entities from a given text corpus;
– integration between available metadata, extracted entities present in the text, and data from external knowledge bases;
– consolidation of the local data in a KG stored in a triple store;
– search and exploration of the corpus through the navigation of the KG in a composite user interface.

In order to ensure the whole solution is useful for potential users it has been implemented and evaluated within a specific case study: exploring the book catalog of a medium-sized publishing house specialized mainly in ancient history. The concrete case study offered the context for fruitful exchange among the stakeholders that are often involved in scenarios of information retrieval and library search: who maintain the corpus (the publisher), who need to search the corpus (researchers of the field and interested individuals), who develop the software solution (in this case the authors of the present study).

The remaining sections are organized as follows. Section 2 presents related work about visual information seeking. Section 3 reports the design process starting from identifying user requirements to the development and implementation of the final interface. Section 4 describes the pipeline of the proposed system and the technologies used during the implementation. Section 5 reports the evaluation process and analyzes the findings. Finally, Section 6 discusses future research directions.

2 Related work

In this section, relevant literature is briefly surveyed for relevant works, starting from traditional systems for visual information seeking to tools for semantic enrichment of unstructured text and visualization/exploration of semantic data as KGs, both in the general case and the specific case of a corpus of books.

2.1 Traditional systems

There has been a large amount of work in the literature about visual information seeking [19,3]. The first attempts to create a visual search interface, have been done in the early 1990s [2], where some researchers had applied direct manipulation principles to search interfaces, creating what they called dynamic queries [1]. These are visual query systems, often based on the query-by-example paradigm[20]: search interfaces where users can manipulate sliders and other graphical controls to change search parameters. The results of those changes are immediately displayed to them in some visualization. A limit of these systems, for unstructured information like books, is that exploring and filtering by basic metadata (i.e., author, title, etc.) can be useful, but it is often insufficient.

2.2 Semantic enrichment

There has hence been recently a lot of research on how to attach semantics to unstructured data [17], through processes like NERL.

The GLOBDEF system [15] works with pluggable enhancement modules, which are dynamically activated to create on-the-fly pipelines for data enhancement. This tool carries out a semantic enrichment challenge but stops there. It does not include the management of the generated metadata, the integration with existing KGs and the visual exploration of these data through a unique visualization.

Apache Stanbol [3] is a set of components able to offer various services for semantic enrichment, visualization of KG and the management of metadata. It is advantageous and can be integrated with our system, but on itself, it does not offer a ready-to-use system.

2.3 Visualization of semantic data

The extracted semantics can then be extremely useful for exploring the data, but they are not fixed and homogeneous like a set of predefined metadata. Therefore, data models and visual user interfaces need to deal with these complex and heterogeneous data. The Semantic Web [4] and Linked Data [6] efforts deal with data modelling, integration, and interaction of this kind of data on the Web. These efforts lately contributed to the emergence of KGs as a way to organize complex data-sets integrating multiple sources [10].

Many user interfaces for visualization and exploration of KGs exist, and new ones are being developed every year, especially in the context of Semantic Web and Linked Data technologies [16,12,5].

Metaphactory [11] is a platform for building KG applications, can be easily integrated into other software infrastructures; for the loading and visualization of RDF KGs, it uses Ontodia [13], which represents one of the most powerful free tools [9] and includes a range of elements that support a variety of interaction techniques. The visualization paradigm is based on the idea of loading in the main panel of the tool the fragment of interest of the entire KG (which can consist of local data, a remote SPARQL endpoint, or the merge of multiple such sources). Entities can be found through textual search and then dragged to the main panel. Connections among entities are shown, and new entities can also be added by expanding the connections of shown entities.

A customized version of Ontodia is included in the software system presented in this paper. The customization enables the use of such KG navigation to search in a text corpus of documents.

2.4 Exploration of a digital library

Many tools face the challenge of the exploration of the contents of a digital library. In particular, two are in the same direction of this work.

Yewno Discover [7] is an integrated system that offers classification and visual exploration of academic materials to help scholars in their research, but

[3] see `https://stanbol.apache.org/`

is not adaptable and flexible to different contexts of use, except with ad hoc adjustments.

Talk to Books[4] is a tool by Google to explore ideas and discover books by getting quotes that respond to user's queries helping users find interesting books that may not be available through keyword search, but does not admit the visual exploration of the RDF KG which allows users to discover, visually, connections between concepts and books.

3 System requirements

The specific use case of the publishing house offered the opportunity to adopt a user-centred design approach to identify and refine the system requirements. From informal interviews with publishing house representatives and a team of researchers in the same domain, an initial set of requirements has been identified:

- the user should be able to search entities textually;
- for an entity, the user should be able to see the relevant books;
- the user should be able to navigate among entities, following semantic relationships between them;
- for a document, the user should be able to access the basic information and be informed on how to obtain it (buy it from a bookstore, borrow it from a library, etc.);
- any user should be able to perform the operations without the need of being taught how to, following established interaction patterns and metaphors.

4 The system

The software system has been implemented and tested in the context of the specific use case, but it is designed for general use. The aim is to offer a ready-to-use package to explore *visually* any corpus of texts through a specialized KG.

4.1 Software modules

The system is composed of a pipeline to build the KG and a web-based front end to search the corpus using the KG. The pipeline is composed of three steps: newly added documents of the corpus enter the pipeline; in the second step, semantic enrichment services extract information from the documents; in the third step, the generated data is consolidated locally in a way that it can also be integrated with additional data provided by external services. RDF is used to represent all the data items in the pipeline, employing existing vocabularies and ontologies whenever possible and creating new terms if needed.

[4] see `https://books.google.com/talktobooks/`

4.2 The user interface

The visual user interface is composed of two main components (see Fig. 1). The first component contains the visualization and search of the entities contained in the KG. It is a customized version of the Ontodia workspace (described in Section 2.3). The second component shows the list of documents associated with the selected entity, offering further interaction with them.

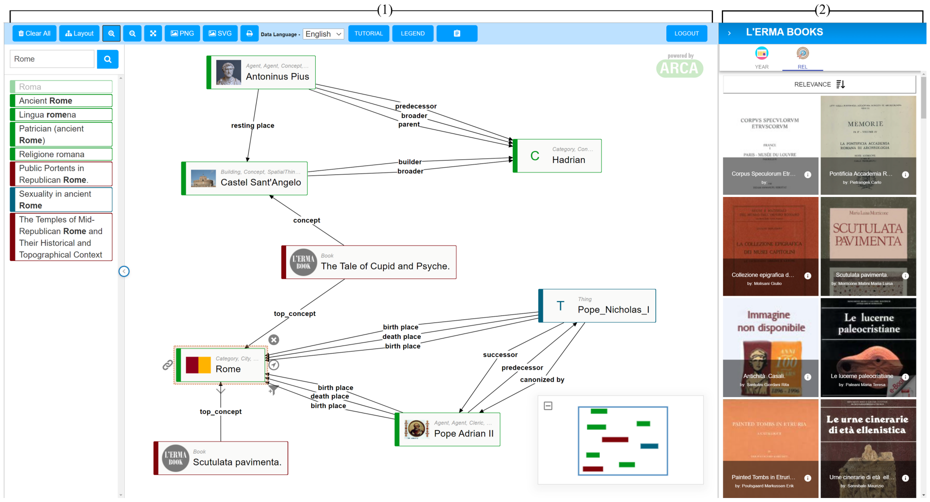


Fig. 1. The user interface.

Exploration of the knowledge graph. The knowledge exploration component (see part 1 of Fig. 1, left side) has the following features.

Searching graph entities. The panel on the left can be used to search for entities in the knowledge graph, corresponding in the use case mainly with entities from DBpedia. For example, typing "Rome" the user gets all the entities containing that string in their label. One or more of the returned entities (e.g., the one corresponding to the city of Rome) may be loaded to the graph navigation panel through drag-and-drop.

Knowledge graph navigation. The central panel allows the user to navigate the KG. Starting from any shown entity, its connections can be expanded (hence adding the connected entities to the graph), either generically showing them all or selecting only some connections of a specific semantic type[5] (e.g., *birth place*). Furthermore, the connections among shown entities are shown by default, as they may be of interest. This panel is coordinated with the document list panel

[5] By the informal term of *semantic connection type* we refer here to RDF properties

(described below and shown in part 2 of Fig. 1, right side) so that the latter shows the list of documents which include, as a topic, the entity currently selected in the former.

Documents as entities. Apart from being shown in the document list panel, documents can also beo explored as entities themselves in the KG exploration. They are linked to their topics by two types of semantic connections: *concept* for any entity found in the text, *top concept* for the ones recognized as main topics for that text. This choice enables different ways to interact with the system:

- starting from a document, exploring its topics and then possibly other documents from them (e.g., in Fig. 1, from the book *The Tale of Cupid and Psyche* to the topic *Rome* and then to the book *Scutulata Pavimenta*);
- from shown entities, visualizing which documents are about two or more of them (e.g., in Fig. 1, the book *The Tale of Cupid and Psyche* is both about *Rome* and, specifically, about *Castel Sant'Angelo*).

Kinds of entities. Different colours are used as an aid to distinguish three broad sets of entities:

- DBpedia entities not found in the corpus are in blue;
- DBpedia entities found at least once in the corpus are in green;
- documents are in red.

Document list. The document list panel (part 2 of Fig. 1, rigth side), which can be shown or hidden as needed, shows the list of documents associated with the entity currently selected in the graph exploration panel, i.e., the documents whose extracted entities include that one. The documents may be shown ordered by year of publication or by relevance (if for that document it is a *main topic* or just a *topic*). By clicking on the *info* button of a book, a modal window with further information on the document is opened. The information includes the list of *snippets*, i.e., all the textual contexts of the document in which the selected concept has been found.

5 Evaluation

As anticipated, the system has been applied to the use case of a publishing house specialized in ancient history. Specifically, it has been tested on a corpus of 112 books, a subset of the full catalog of the editor. The evaluation has been carried out with the help of a group of researchers who are experts of ancient history and specifically of the topics covered by the set of books. The evaluation of the system has been divided into three phases. In the first phase, the researchers, as domain experts, assessed the quality of the semantic enrichment process. In the following two phases, the system as a whole has been evaluated through user tests. In both phases, users interact with the visual user interface, in the first phase without given constraints, in the second phase with a set of tasks and a more structured setup.

5.1 Quality of entity extraction

The researchers have read deeply ten books contained in the system, and for each book have evaluated the quality of:

- the correspondence of the *top concepts* with the main topics;
- the correspondence of the *concepts* with entities mentioned in the book;
- the disambiguation of the words.

The findings were analyzed qualitatively, as the purpose was to test the feasibility of the whole system rather than to evaluate the NERL method per se. The quality has been deemed sufficient to be used effectively. Nevertheless, some issues emerged. They are described in section 5.4 along with ideas to approach them.

5.2 Direct observation of unconstrained navigation

In the second phase of the evaluation, two users were invited to experiment using the user interface without any specific constraints.

Three types of reactions have been observed:

- positive surprise in finding and verifying relationships between concepts that they were already aware of;
- amazement in finding new unknown relationships;
- displeasure in not finding expected relationships, due to the lack of content.

Overall, from this first observation of the user interface use, the results have been that the navigation was smooth and stimulating in the exploration of the contents.

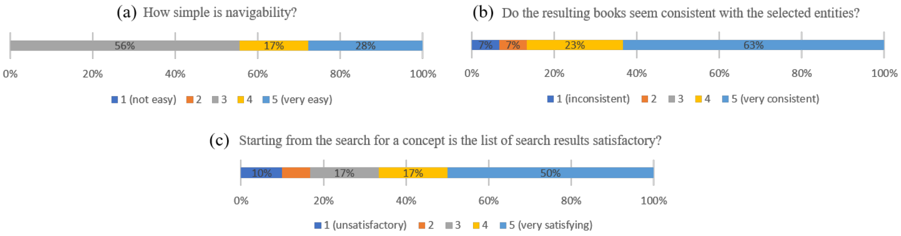
5.3 Task-driven usage and questionnaire

In the second part of the evaluation, six users participated in a task-oriented evaluation.

Before starting with the compilation of the questionnaire, users were invited to watch four video tutorials in the appropriate section of the interface to familiarize themselves with the ARCA commands and functions. They were given a questionnaire containing task-instructions interspersed with the related questions, some of them open ended, some of them requiring a number on a Likert scale from 1 to 5. The questionnaire was divided into four sections: (I) three basic tasks followed each one by a numeric question on the difficulty; (II) five identical macro-tasks (decomposed in seven sub-tasks) each one of them consisting in a set of search-exploration steps starting from a different topic, chosen by the user, interspersed by two questions after each sub-task, one numeric and one open ended, to evaluate the data quality of the explored results; (III) four general open ended questions; (IV) other three pairs of questions, one numeric and one open ended, to evaluate the technical aspects of the interface.

Using and aggregating Likert scale scores was helpful in quantitatively summarizing the sentiment of the users, albeit the number of users involved in the test is too small to look for statistical relevance.

I To answer the first section of the questionnaire, users were assigned three precise activities to perform, and on the basis of the results obtained, it been assessed the ease of use of the system and the satisfaction obtained by the exploration of resources. From Figure 2a it can be seen that 100% of the test considered the navigability of ARCA from simple to very simple. In Figure 2b the 87% consider consistent the books related to the selected entities and in Figure 2c 83% is satisfied with the list of search results connected to a concept.



(a) How simple is navigability?

(b) Do the resulting books seem consistent with the selected entities?

(c) Starting from the search for a concept is the list of search results satisfactory?

Fig. 2. Answer distribution - Sec. I of the questionnaire

II In the second section of the questionnaire, instead of assigning specific tasks, each of the six users was given the freedom to choose five different concepts from which to start their own research and five different books to explore at own choice. For each chosen concept and book, the questions assessed the quality of the explored search results. From this group of questions, it emerged that for the users, the obtained results are mostly coherent with what they expect.

III In the third section, four general questions were asked, which are shown below along with an outline of user responses.

What are the most useful features of ARCA?
Among the most useful features, users that have tested the interface have identified:

- the possibility of identifying and exploring links between concepts which makes the research method immersive and stimulating;
- the search for books starting from concepts and, above all, starting from other books;
- the fact that it is an open system and modulated on improvable and replaceable components;
- it is a potentially very large container of editorial products;
- the ability to save search paths and export them as SVG and PNG images;
- the display of integrated data (texts, images, links);

- the direct connection with the catalogs of the publishing house;
- the possibility of verifying the concepts extracted from the corpus of books through the "info" button of the component containing the books;
- the division of the concepts extracted from books into concepts and top concepts.

What are ARCA weaknesses?
Among the weaknesses of ARCA at the actual stage, users have identified:

- sometimes errors of disambiguation and restitution of entities not always relevant;
- sometimes errors in extracting concepts from the text;
- little information contained in the system to satisfy curiosity in exploring a larger catalog of books to discover more connections;
- the system offers multiple features that must be properly explained to allow the user a complete browsing experience.

What features do you think is useful to add to ARCA to make it a better system?
The features that, for the users, could improve the system are:

- integration with other KGs;
- improvement of the entity extraction module;
- improve the order and organization of the connected elements to the entities;

Add any other notes and thoughts useful to improve ARCA.
Users have noted that it might be interesting to implement the following features:

- offer the possibility to save the search history;
- create personal bibliographic lists with the results of the searches in a dedicated space on the board (exportable and downloadable);
- create a support system that can help the users to know in a more easily and understandable way, how to take full advantage of all the features of ARCA

IV In the fourth section, there are questions to assess the technical aspects of the interface. The Likert scale questions and answers reveal that the loading times and the organization of the visualization of the results can be improved. In the motivation following the Likert scale questions, users told that they have experienced slowdowns in particular in viewing the book catalog, and in the loading of video tutorials.

5.4 Discussions and limitations

The system obtained a more than satisfactory performance in recommending relevant editorial products and received a high score in terms of usability; the simplicity of use; user satisfaction with the results shown; consistency of the contents with the s domain of the publishing house; attractiveness of the system.

Nonetheless, the evaluation of the quality of the entity extraction (Section 5.1) highlighted some issues that need to be addressed in future versions and considered in related works, like the wrong resolution of acronyms and abbreviations and the incorrect disambiguation of entities due to the absence of the correct entity in the DBpedia KG.

Another feedback is that some users complained about a relatively small amount of information in the internal KG (built with the concepts of 112 books and the related metadata). It is expected that when the catalog of books is numerically more significant, the chance of discovering new information and connections while browsing the KG will increase.

Finally, based on initial observations, it has been seen that using the system at first glance can be difficult without viewing the video tutorials. As a lighter alternative to video tutorials, an help component could be implemented to accompany the users in the first searches and thus make them independent in exploiting all the possibilities of exploration that the system offers.

6 Conclusions and future work

ARCA is an innovative system based on the visual semantic search that allows the exploration of a text corpus. Through a knowledge graph-based navigation, users can start from any relevant entity and reach other entities related to it, discovering in which books or articles each entity is present and evaluating which of these results are useful for their research. The user studies conducted so far confirm the amenability of the proposed system to domain experts who were able to perform non-trivial tasks of search and exploration, tasks that would be more cumbersome to execute with the search tools they are used to. Feedback gathered from users suggest that the proposed exploration mechanism tends to amplify the user experience by also offering opportunities for further study and discovery of sources, themes, and materials, which have the potential of enriching the research process with new ideas.

Through the presented user study, many desiderata have been collected, and they can be used to guide further development and experimentation in this context. Furthermore, in order to extend the evaluation to more users, a more extensive indirect observation study has been planned. In addition to the questionnaire, the analysis will be further completed with objective data gathered by tracking users' activity through interaction logs.

As a potential future direction of research and development, the scope of the semantic enrichment process could be broadened to other document elements, such as images and captions, enriching KG exploration.

References

1. Ahlberg, C., Shneiderman, B.: Visual information seeking: Tight coupling of dynamic query filters with starfield displays. In: The Craft of Information Visualization, pp. 7–13. Elsevier (2003)
2. Ahlberg, C., Williamson, C., Shneiderman, B.: Dynamic queries for information exploration: An implementation and evaluation. p. 619–626. CHI '92, Association for Computing Machinery, New York, NY, USA (1992)
3. Baeza-Yates, R., Ribeiro-Neto, B., et al.: Modern information retrieval, vol. 463. ACM press New York (1999)
4. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American **284**(5), 34–43 (2001)
5. Bikakis, N., Sellis, T.: Exploration and visualization in the web of big linked data: A survey of the state of the art. arXiv preprint arXiv:1601.08059 (2016)
6. Bizer, C., Heath, T., Berners-Lee, T.: Linked data-the story so far. Int. J. on Semantic Web and Information Systems **5**(3), 1–22 (2009)
7. Bolina, M.: Yewno Discover. Nordic Journal of Information Literacy in Higher Education **11**(1) (2019)
8. Ceriani, M., Bernasconi, E., Mecella, M.: A streamlined pipeline to enable the semantic exploration of a bookstore. In: Digital Libraries: The Era of Big Data and Data Science - 16th Italian Research Conference on Digital Libraries, IRCDL 2020, Bari, Italy, January 30-31, 2020, Proceedings. Communications in Computer and Information Science, vol. 1177, pp. 75–81. Springer (2020)
9. Dudás, M., Lohmann, S., Svátek, V., Pavlov, D.: Ontology visualization methods and tools: a survey of the state of the art. Knowledge Eng. Review **33**, e10 (2018)
10. Ehrlinger, L., Wöß, W.: Towards a definition of knowledge graphs. In: SEMANTiCS (2016)
11. Haase, P., Herzig, D., Kozlov, A., Nikolov, A., Trame, J.: metaphactory: A platform for knowledge graph management. Semantic Web **10**, 1–17 (06 2019)
12. Marie, N., Gandon, F.L.: Survey of linked data based exploration systems. In: IESD@ISWC (2014)
13. Mouromtsev, D., Pavlov, D., Emelyanov, Y., Morozov, A., Razdyakonov, D., Galkin, M.: The simple web-based tool for visualization and sharing of semantic data and ontologies. In: International Semantic Web Conference (2015)
14. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Investigationes **30**(1), 3–26 (2007)
15. Nisheva-Pavlova, M., Alexandrov, A.: GLOBDEF: A Framework for Dynamic Pipelines of Semantic Data Enrichment Tools. In: Proc. of MTSR 2018. pp. 159–168. Springer (2018)
16. Po, L., Bikakis, N., Desimoni, F., Papastefanatos, G.: Linked data visualization: Techniques, tools, and big data (2020)
17. Ristoski, P., Paulheim, H.: Semantic web in data mining and knowledge discovery: A comprehensive survey. Journal of Web Semantics **36**, 1–22 (2016)
18. Shen, W., Wang, J., Han, J.: Entity linking with a knowledge base: Issues, techniques, and solutions. IEEE Transactions on Knowledge and Data Engineering **27**(2), 443–460 (2014)
19. Shneiderman, B.: The eyes have it: A task by data type taxonomy for information visualizations. In: The craft of information visualization, pp. 364–371. Elsevier (2003)
20. Zloof, M.M.: Query-by-example: A data base language. IBM Syst. J. **16**, 324–343 (1977)